Overlaying a hypergraph with a graph with bounded maximum degree

Frédéric Havet, Dorian Mazauric, Viet-Ha Nguyen, Rémi Watrigant

Recouvrement d'un hypergraphe par un graphe de degré borné pour déterminer les contacts d'un assemblage macromoléculaire

Frédéric Havet[1,2,3] et Dorian Mazauric[1,3] et Viet-Ha Nguyen[1,2,3] et Rémi Watrigant[4]

[1] *Université Côte d'Azur, France*
[2] *CNRS, I3S, France*
[3] *Inria Sophia Antipolis Méditerranée, France*
[4] *Université de Lyon, France*

Soient G un graphe et H un hypergraphe avec le même ensemble de sommets et soit F un graphe fixé. Le graphe G F*-revouvre* une hyperarête S de H si F est un sous-graphe couvrant du sous-graphe de G induit par S. Le graphe G F*-revouvre* H s'il F-revouvre chaque hyperarête de H.

Nous avons analysé la complexité des deux problèmes suivants. Le premier problème, $(\Delta \leq k)$ F-OVERLAY, consiste à décider s'il existe un graphe de degré maximum au plus k qui F-revouvre un hypergraphe H donné. C'est un cas particulier du second problème, MAX $(\Delta \leq k)$ F-OVERLAY, qui, étant donné un hypergraphe H et un entier s, consiste à décider s'il existe un graphe de degré maximum au plus k qui F-revouvre au moins s hyperarêtes de H.

Nous prouvons une dichotomie polynomial/$\mathcal{NP}$-complet pour le problème MAX $(\Delta \leq k)$-F-OVERLAY dépendant de la paire (F,k), et prouvons la complexité du problème $(\Delta \leq k)$ F-OVERLAY pour un grand nombre de paires (F,k).

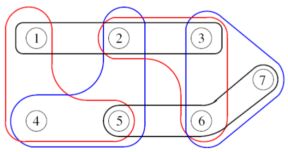
Ces deux problèmes modélisent un problème central en biologie structurale computationnelle : la détermination des contacts entre les protéines d'un assemblage macromoléculaire. Les sommets sont les protéines, les hyperarêtes sont les complexes connus, le graphe F est le *graphe générique* dont les arêtes correspondent aux contacts entre les protéines de l'assemblage. Déterminer le graphe G revient alors à trouver les contacts entre les protéines de telle sorte que le graphe F soit un sous-graphe couvrant dans chaque hyperarête et de telle sorte que le degré soit borné (une protéine est en contact avec un nombre limité d'autres protéines). Enfin, ces problèmes sont d'intérêt plus général pour les problèmes d'inférence de réseaux.

Mots-clefs : hypergraphes, graphes, algorithmes, complexité, biologie structurale computationnelle

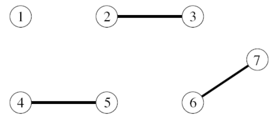
1 Introduction

1.1 Context and definition of the problems

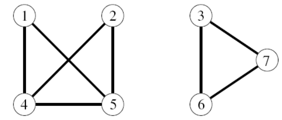
A major problem in structural biology is the characterization of low resolution structures of macromolecular assemblies [BDD$^+$12, SR07]. To attack this very difficult question, one has to determine the plausible contacts between the subunits (e.g. proteins) of an assembly, given the lists of subunits involved in all the complexes. We assume that the composition, in terms of individual subunits, of selected complexes is known. Indeed, a given assembly can be chemically split into complexes by manipulating chemical conditions. This problem can be conveniently modeled by graphs and *hypergraphs* where a hypergraph H has a set of vertices $V(H)$ and a collection of subset of vertices, called hyperedges, $E(H)$. So, vertices represent the subunits and whose hyperedges are the complexes. An example of hypergraph H is in Figure 1a, it has seven vertices and six hyperedges (each one is a subset of three vertices). We are then looking for a graph G with the same vertex set as H whose edges represent the contacts between subunits, and satisfying (i) some local properties for every complex (*i.e.* hyperedge), and (ii) some other global properties. In our problems, we use the following notion of *overlayment*.



(a) A hypergraph H



(b) A graph G_1 O_3-overlaying H



(c) A graph G_2 C_3-overlaying three hyperedges of H

FIGURE 1: An example of two problems on hypergraph H. In the figure 1b, each hypergraph (*e.g.* $S_1 = \{1,4,5\}$) is overlaid by a graph with three vertices and one edge (e.g. the edge in $G[S_1]$ is (4,5). In the figure 1c, G_2 overlays three hyperedges S_1, $S_2 = \{2,4,5\}$, $S_3 = \{3,6,7\}$, each induced subgraph $G[S_i]$ (for $i = \{1,2,3\}$) is a C_3 graph.

Definition 1 A graph G $\mathcal{F}$*-overlays* a hyperedge S if there exists $F \in \mathcal{F}$ such that F is a spanning subgraph of $G[S]$, and it $\mathcal{F}$*-overlays* H if it $\mathcal{F}$-overlays every hyperedge of H.

As said previously, the graph sought will also have to satisfy some global constraints. Since in a macromolecular assembly the number of contacts is small, the first natural idea is to look for a graph G with the minimum number of edges. This led to the MIN-$\mathcal{F}$-OVERLAY problem : given a hypergraph H and an integer m, decide if there exists a graph G $\mathcal{F}$-overlaying H such that $|E(G)| \leq m$. A typical example of a family $\mathcal{F}$ is the set of all connected graphs. Agarwal et al. [ACCC15] focused on MIN-$\mathcal{M}(\mathcal{F})$-OVERLAY for this particular family in the aforementioned context of structural biology. However, this problem was previously studied by several communities in other domains, for example, in the design of vacuum systems [DK95], scalable overlay networks [CMTV07], and reconfigurable interconnection networks [FHWE08].

In this paper, we consider the variant in which the additional constraint is that G must have a bounded maximum degree : this constraint is motivated by the context of structural biology, since a subunit (e.g. a protein) cannot be connected to many other subunits. Moreover, we mainly consider the case when the family $\mathcal{F}$ contains a unique graph F. This yields the following problem for any graph F and an integer k.

> $(\Delta \leq k)$-F-OVERLAY
> Input : A hypergraph H.
> Question : Does there exist a graph G F-overlaying H such that $\Delta(G) \leq k$?

A natural generalization of this problem is to find a graph with maximum degree at most k which overlays a maximum number of hyperedges.

> MAX $(\Delta \leq k)$-F-OVERLAY
> Input : A hypergraph H and a positive integer s.
> Question : Does there exist a graph G F-overlaying at least s hyperedges in $E(H)$ such that $\Delta(G) \leq k$?

Examples of the two problems on two graphs O_3 (which has three vertices and one edge) ; and C_3 (the complete graph on 3 vertices) which are showed in Figure 1 : in both problems we are given the hypergraph H (Figure 1a). In the first one, one checks easily that the graph G_1 (Figure 1b) O_3-overlays all hyperedges of H and $\Delta(G_1) = 1$. Then G_1 is an output of $(\Delta \leq 1)$-O_3-OVERLAY. In the second example, graph G_2 (Figure 1c) C_3-overlays three hyperedges. Graph G_2 is a solution of MAX $(\Delta \leq 3)$-C_3-OVERLAY with $s = 3$.

Observe that there is an obvious reduction from $(\Delta \leq k)$-$\mathcal{F}$-OVERLAY to MAX $(\Delta \leq k)$-$\mathcal{F}$-OVERLAY (by setting $s = |E(H)|$).

By definition those two problems really make sense only for $|F|$-uniform hypergraphs *i.e.* hypergraphs whose hyperedges are of size $|F|$. Therefore, we always assume the hypergraph to be $|F|$-uniform.

1.2 Our contributions

We study the dichotomy of complexity ($\mathcal{P}$ vs $\mathcal{NP}$-complete) of the problems $(\Delta \leq k)$-F-OVERLAY and MAX $(\Delta \leq k)$-F-OVERLAY. We have the following observations. First, if F is a graph with maximum degree greater than k, then solving $(\Delta \leq k)$-F-OVERLAY or MAX $(\Delta \leq k)$-F-OVERLAY is trivial as the answer is always 'No'. Thus, we only study the problems when $\Delta(F) \leq k$. Second, if F is an empty graph, then MAX $(\Delta \leq k)$-F-OVERLAY is also trivial, because for any hypergraph H, the empty graph on $V(H)$ vertices F-overlays H. Hence the first natural interesting cases are the graphs with one edge. For every integer $p \geq 2$, we denote by O_p the graph with p vertices and one edge. We obtain the following result.

Theorem 1 *Let $k \geq 1$ and $p \geq 2$ be integers. If $p = 2$ or if $k = 1$ and $p = 3$, then* MAX $(\Delta \leq k)$-O_p-OVERLAY *and* $(\Delta \leq k)$-O_p-OVERLAY *are polynomial-time solvable. Otherwise, they are $\mathcal{NP}$-complete.*

For general graphs, we also obtain the complete dichotomy results of the problem MAX $(\Delta \leq k)$-F-OVERLAY.

Theorem 2 MAX $(\Delta \leq k)$-F-OVERLAY *is polynomial-time solvable if either $\Delta(F) > k$, or F is an empty graph, or $F = O_2$, or $k = 1$ and $F = O_3$. Otherwise it is $\mathcal{NP}$-complete.*

In addition, we investigate the complexity of $(\Delta \leq k)$-F-OVERLAY problems. We believe that each such problem is either polynomial-time solvable or $\mathcal{NP}$-complete. However the dichotomy seems to be more complicated than the one for MAX $(\Delta \leq k)$-F-OVERLAY. We exhibit several pairs (F, k) such that $(\Delta \leq k)$-F-OVERLAY is polynomial-time solvable, while MAX $(\Delta \leq k)$-F-OVERLAY is $\mathcal{NP}$-complete. Precisely, the cases we look at are in $\mathcal{P}$ are listed as follows.

— $(\Delta \leq k)$-F-OVERLAY when F is a connected k-regular graph.

— $(\Delta \leq 2)$-$\mathcal{P}$-OVERLAY where $\mathcal{P}$ is the family of paths.

— $(\Delta \leq 3)$-C_4-OVERLAY.

Due to space constraints, several proofs are omitted. The full version [HMNW19] of the paper is available on hal (hal-02025469). In the next section, we expose the main ideas to obtain polynomial-time algorithms when $\mathcal{F}$ is $\{O_2\}$ or $\{O_3\}$ (Theorem 1) or the family of paths.

2 Some polynomial-time algorithms (sketch)

Most notations are standard. We denote by $K(H)$, the graph obtained by replacing each hyperedge by a complete graph. The *edge-weight function induced by H* on $K(H)$, denoted by w_H, is defined by $w_H(e) = |\{S \in E(H) \mid e \subseteq S\}|$.

2.1 Polynomial cases of graphs O_p

Consider first the case when $p = 2$. Let H be a 2-uniform hypergraph. Every hyperedge is an edge, so $K(H) = H$. Moreover, a (hyper)edge of H is O_2-overlaid by G if and only if it is in $E(G)$. Hence MAX $(\Delta \leq k)$-O_2-OVERLAY is equivalent to finding a maximum k-matching (that is a subgraph with maximum degree at most k) in $K(H)$. This problem is polynomial-time solvable, see [Sch03, Chap. 31].

Now consider the case $p = 3$. Since the edge set of a graph with maximum degree 1 is a matching, and every hyperedge overlaid by O_3 contains an edge of a matching, thus MAX $(\Delta \leq 1)$-O_3-OVERLAY is equivalent to finding a maximum-weight matching in the edge-weighted graph $(K(H), w_H)$. This can be done in polynomial-time, see [Jun13, Chap. 14].

Hence, we have the following : if $p = 2$ or ($p = 3$ and $k = 1$), then $(\Delta \leq k)$-F-OVERLAY and MAX $(\Delta \leq k)$-F-OVERLAY are polynomial-time solvable.

2.2 Polynomial case of paths

In this section, we consider the family of paths $\mathcal{P}$ instead of only one precise path P. Then, a given hypergraph is not necessary to be $|P|$-uniform. We will give briefly the proof of the following theorem.

Theorem 3 $(\Delta \leq 2)$-$\mathcal{P}$-OVERLAY *where $\mathcal{P}$ is the family of paths.*

Clearly, if H is not connected, it suffices to solve the problem on each of the components and to return 'No' if the answer is negative for at least one of the components, and 'Yes' otherwise. Henceforth, we shall now assume that H is connected.

We first define an *intersection graph* of a hypergraph H, denoted by $IG(H)$ as follows : a vertex of $IG(H)$ is a hyperedge, and two vertices are adjacent if the two corresponding hyperedges intersect.

Let $\mathbb{C}_l$ be the circle of circumference l. We identify the points of $\mathbb{C}_l$ with the integer numbers (points) of the segment $[0,l]$, (with 0 identified with l). Let two function l_H, s_H where $l_H(S) = |S| - 1$ for any $S \in E(H)$, and $s_H(S,S') = |S \cap S'| - 1$ for all $S, S' \in E(H)$. A *circular-arc graph* is the intersection graph of a set of arcs on $\mathbb{C}_l$. A set $\mathcal{A}$ of arcs such that $IG(\mathcal{A}) = G$ is called an *arc representation* of G. We denote by A_v the arc corresponding to v in $\mathcal{A}$. Let G be a graph and let $l : V(G) \to \mathbb{N}$ and $s : E(G) \to \mathbb{N}$ be two functions. An arc representation $\mathcal{A}$ of G is *l-respecting* if A_v has length $l(v)$ for any $v \in V(G)$, *s-respecting* if $A_v \cap A_u$ has length $s(u,v)$ for all $uv \in E(G)$, and (l,s)*-respecting* if it is both l-respecting and s-respecting.
Then, we obtain the following result.

Claim 1 *Let H be a connected hypergraph on n vertices. There is a graph $\mathcal{P}$-overlaying H if and only if $IG(H)$ admits an (l_H, s_H)-respecting arc representation into $\mathbb{C}_n$.*

One can easily adapt the algorithm given by Köbler et al. [KKW15] for (l,s)-respecting interval representations to decide in linear time whether a graph admits an (l,s)-respecting arc representation in $\mathbb{C}_n$ for every integer n.

3 Further research

A first open question is to close the dichotomy of $(\Delta \leq k)$-F-OVERLAY, that is characterize the pairs (F,k) for which $(\Delta \leq k)$-F-OVERLAY is polynomial-time solvable and those for which it is $\mathcal{NP}$-complete.

Theorem 2 characterizes the complexity of MAX $(\Delta \leq k)$-$\mathcal{F}$-OVERLAY when $\mathcal{F}$ contains a unique graph. It would be nice to extend this characterization to families $\mathcal{F}$ of arbitrary size, that is characterize the pairs $(\mathcal{F},k)$ for which MAX $(\Delta \leq k)$-$\mathcal{F}$-OVERLAY is polynomial-time solvable and those for which it is $\mathcal{NP}$-complete.

Other interesting open questions are detailed in [HMNW19].

Références

[ACCC15] Deepesh Agarwal, Christelle Caillouet, David Coudert, and Frédéric Cazals. Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems. *Molecular and Cellular Proteomics*, 14 :2274–2284, April 2015.

[BDD+12] Rebecca Burnley, Eugen Damoc, Eduard Denisov, Alexander Makarov, and Albert Heck. High-sensitivity orbitrap mass analysis of intact macromolecular assemblies. *Nature methods*, 9, 10 2012.

[CMTV07] Gregory Chockler, Roie Melamed, Yoav Tock, and Roman Vitenberg. Constructing scalable overlays for pub-sub with many topics. In *PODC '07*, pages 109–118. ACM, 2007.

[DK95] Ding Zhu Du and Dean F. Kelley. On complexity of subset interconnection designs. *Journal of Global Optimization*, 6(2) :193–205, 1995.

[FHWE08] Hongbing Fan, Christian Hundt, Yu-Liang Wu, and Jason Ernst. Algorithms and implementation for interconnection graph problem. In *Combinatorial Optimization and Applications*, pages 201–210, 2008.

[HMNW19] Frédéric Havet, Dorian Mazauric, Viet-Ha Nguyen, and Rémi Watrigant. Overlaying a hypergraph with a graph with bounded maximum degree. Research Report RR-9258, Inria Sophia Antipolis, February 2019.

[Jun13] D. Jungnickel. *Graphs, Networks and Algorithms*, volume 5 of *Algorithms and Computation in Mathematics*. Springer-Verlag Berlin Heidelberg, 2013.

[KKW15] Johannes Köbler, Sebastian Kuhnert, and Osamu Watanabe. Interval graph representation with given interval and intersection lengths. *J. Discrete Algorithms*, 34 :108–117, 2015.

[Sch03] Alexander Schrijver. *Combinatorial optimization. Polyhedra and efficiency. Vol. A*, volume 24 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2003.

[SR07] Michal Sharon and Carol V. Robinson. The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annual Review of Biochemistry*, 76(1) :167–193, 2007. PMID : 17328674.